MCMC Methods

Varga Katalin (Ph.D.)

BUTE

2016 Autumn

Gibbs Sampling

Gibbs Sampling

- Easily implemented methods for sampling from multivariate distributions, $p(\theta_1, \dots, \theta_k)$, which are typically a posterior distributions.

- Requirements: easily sampled full conditional posteriors:

    - $p(\theta_1 | \theta_2, \theta_3, \dots, \theta_k)$
    - $p(\theta_2 | \theta_1, \theta_3, \dots, \theta_k)$
    - …
    - $p(\theta_k | \theta_1, \theta_2, \dots, \theta_{k-1})$

Gibbs Sampling Algorithm

- A:

Choose initial values ${\theta_2}^{(0)}, {\theta_3}^{(0)}, \dots, {\theta_n}^{(0)}$.

- B:
  - B1: Draw ${\theta_1}^{(1)}$ from $p\left(\theta_1 \middle| {\theta_2}^{(0)}, {\theta_3}^{(0)}, \dots, {\theta_n}^{(0)}\right)$
  - B2: Draw ${\theta_2}^{(1)}$ from $p\left(\theta_2 \middle| {\theta_1}^{(1)}, {\theta_3}^{(0)}, \dots, {\theta_n}^{(0)}\right)$ eloszlásból
  - Bn: Draw ${\theta_n}^{(1)}$ from $p\left(\theta_n \middle| {\theta_1}^{(1)}, {\theta_2}^{(1)}, \dots, {\theta_{n-1}}^{(1)}\right)$

- C:

Repeat step B N times.

Gibbs Sampling Algorithm

- The $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$Gibbs draws are dependent, but arithmetic means converge to expected values:

$$\frac{1}{N}\sum_{t=1}^{N} \theta_j{}^{(t)} \to E(\theta_j)$$

$$\frac{1}{N}\sum_{t=1}^{N} g(\theta_j{}^{(t)}) \to E[g(\theta_j)]$$

- More generally, $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$ Gibbs sequence converge in distribution to the target posterior $P(\theta_1, \dots, \theta_k)$.

- $\theta_j{}^{(1)}, \theta_j{}^{(2)}, \dots, \theta_j{}^{(N)}$ converge to the marginal distribution of θ_j,
  to $P(\theta_j)$.

Examples of Gibbs Sampling

- Bivariate normal:
  - Joint distribution:

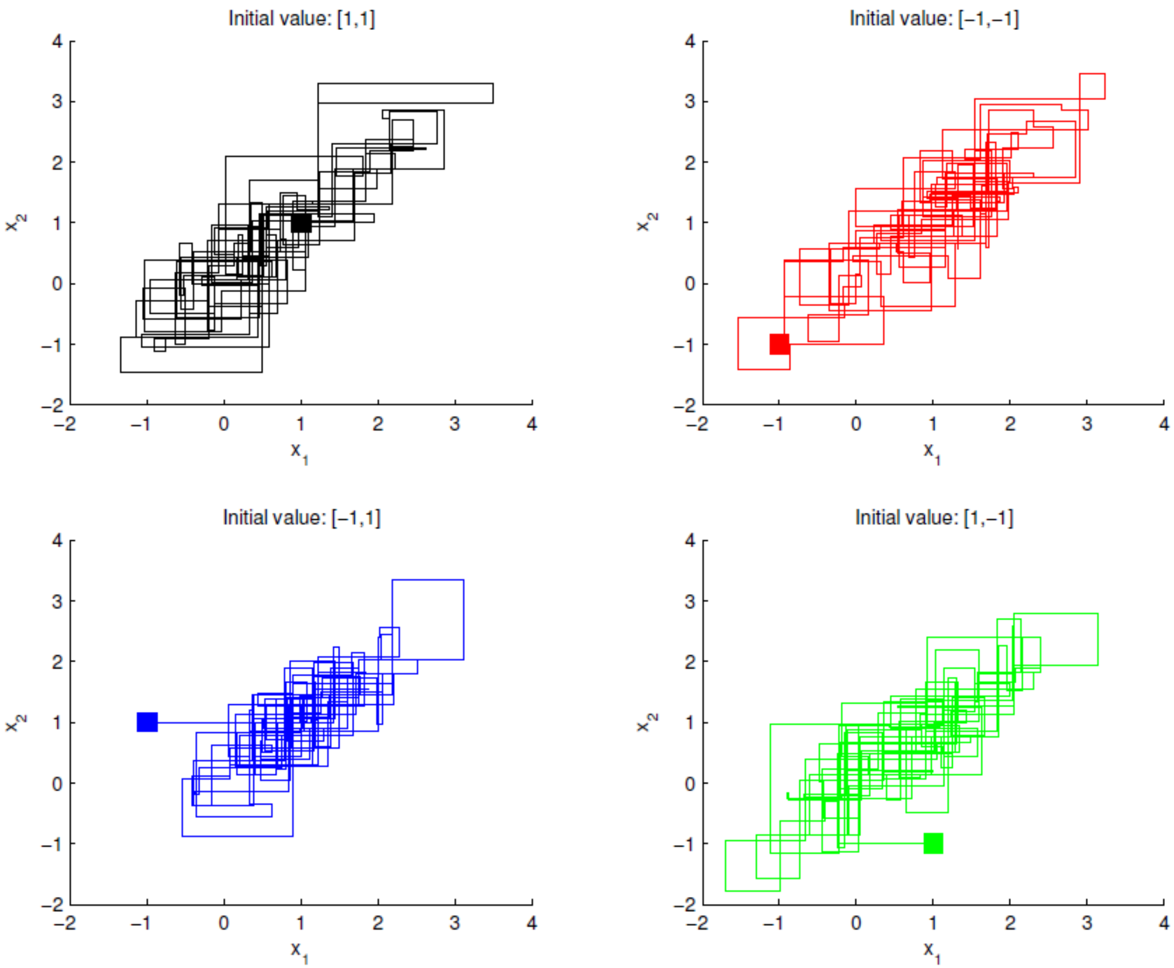
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

  - Full conditional posteriors:

$$\theta_1 | \theta_2 \sim N[\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2]$$

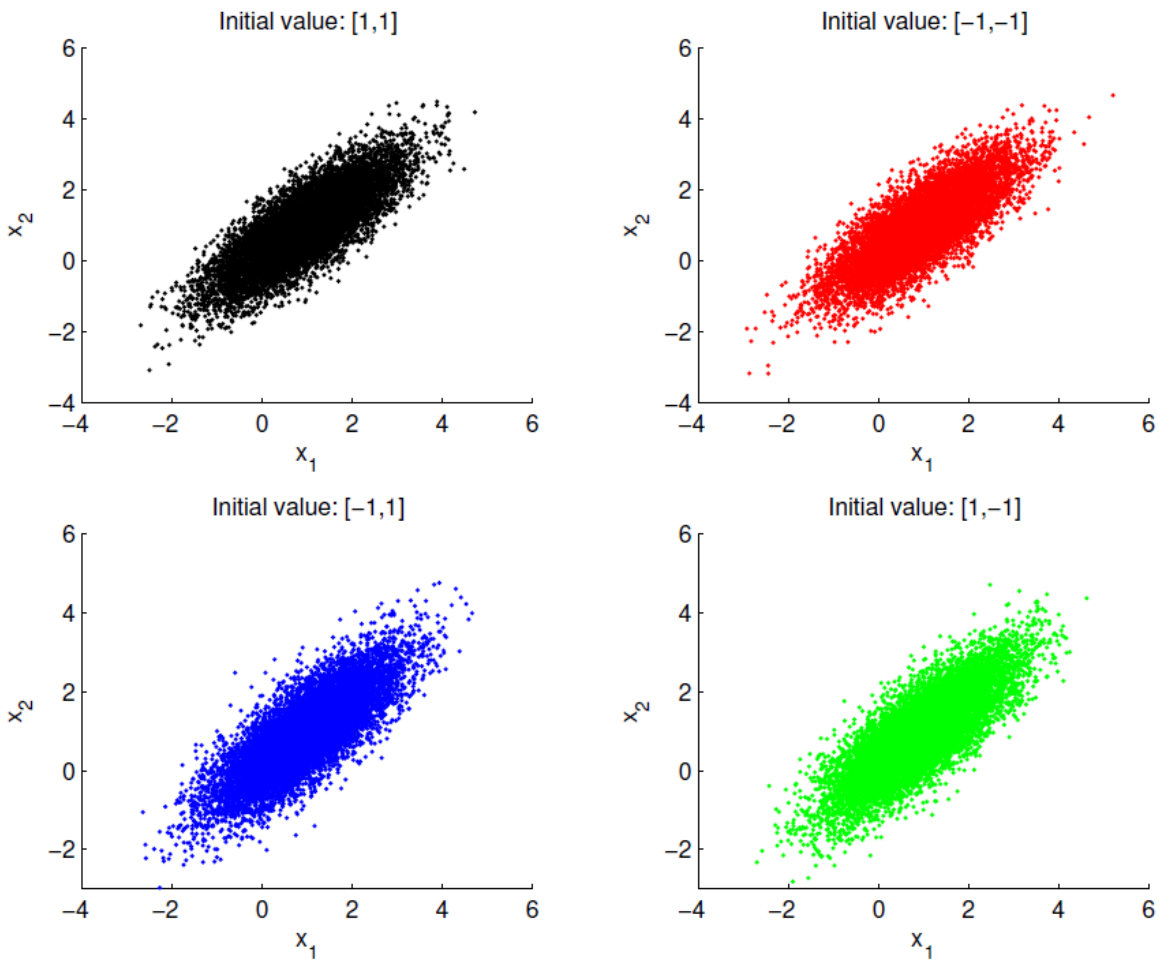
$$\theta_2 | \theta_1 \sim N[\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2]$$

Bivariate Normal – Initial Values don't Matter



Reference: Mattias Villani (Sveriges Riksbank)

Bivariate Normal – Initial Values don't Matter



Reference: Mattias Villani (Sveriges Riksbank)

Metropolis Algorithm

A. Initialize with $\theta = \theta_0$

B. For $t = 1{,}2$, sample a proposal draw:

$$\theta^* | \theta^{(t-1)} \sim q_t\left(\theta^* \middle| \theta^{(t-1)}\right)$$

C. Accept θ^* with probability:

$$r\left(\theta^{(t-1)} \to \theta^*\right) = min\left[\frac{P(\theta^*|y)}{P\left(\theta^{(t-1)}\middle|y\right)}, 1\right]$$

D. If proposal is accepted, set $\theta^{(t)} = \theta^*$, otherwise set $\theta^{(t)} = \theta^{t-1}$.

- Every proposal θ^*that lies uphill is always accepted, downhill moves θ^* accepted with probability $r\left(\theta^{(t-1)} \to \theta^*\right)$.
- It is enough if we can compute the **unnormalized** posterior density, $P(y|\theta)P(\theta)$ for any θ.
- $q_t\left(\theta^* \middle| \theta^{(t-1)}\right)$ must be symmetric i.e. $q_t(\theta_a|\theta_b) = q_t(\theta_b|\theta_a)$.

Metropolis - Choosing the Proposal Distribution

- Common choice of proposal distribution:

$$q_t\left(\theta^*\middle|\theta^{(t-1)}\right) = \mathrm{N}\left[\theta^{(t-1)}, c^2 J^{-1}(\hat{\theta})\right]$$

- Where c is a tuning constant and J is the observed Fisher information:

$$J_y = -\frac{\partial^2 lnP(\theta|y)}{\partial\theta^2}|\theta = \hat{\theta}$$

- A good proposal, $q_t\left(\theta^*\middle|\theta^{(t-1)}\right)$ should have the following properties:
  - Easy to sample
  - Easy to compute $r\left(\theta^{(t-1)} \rightarrow \theta^*\right)$
  - Takes reasonably large jumps in the parameter space
  - Jumps are rejected not too frequently
  - Set c to that average acceptance probability is shomewhere between 0,2 and 0,4

Practical Implementation of MCMC Algorithms

The **autocorrelation** in the simulated sequence $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}$ makes it showmewhat problematic to define the effective number of simulation draws.

- **Inefficiency factor:**

$$IF = 1 + 2\sum_{i=1}^{\infty} \rho_i$$

- Where ρ_i is the autocorrelation at lag i.

- **Effective sample size:**

$$ESS = \frac{N}{IF}$$

- When do we stop sampling?
- How many burn-in iterations to discard?
- Several short sequences or a single long sequence? To thin out or not to thin out?
- Convergence diagnostics.

The Metropolis-Hastings Algorithm

- Generalization of the Metropolis algorithm to **non-symmetric** proposals.
- The acceptance probability is slightly more complicated:

$$r\left(\theta^{(t-1)} \to \theta^*\right) = min\left[\frac{P(\theta^*|y)/q_t\left(\theta^*\middle|\theta^{(t-1)}\right)}{P\left(\theta^{(t-1)}\middle|y\right)/q_t\left(\theta^{\,(t-1)}\middle|\theta^*\right)}, 1\right]$$

- Gibbs sampling is a special case of the MH algorithm where the proposal is the full conditional posterior and every draw is accepted.

- **Independence MH**: $q_t\left(\theta^*\middle|\theta^{(t-1)}\right) = q_t(\theta^*)$. Example: $\theta^* \sim \mathrm{N}\left[\hat{\theta}, J^{-1}\left(\hat{\theta}\right)\right]$.

- Metropolis-Hastings-within-Gibbs:
  - $P(\theta_1|\theta_2, y)$ is an easily sampled distribution
  - $P(\theta_2|\theta_1, y)$ is not easily sampled. MH updating step.

Bayesian Model Inference

- Comparing two models $P_1(x|\theta_1)$ and $P_2(x|\theta_2)$ would be easy if θ_1 and θ_2 were known. Usually they aren't.
- Bayes: average with respect to the prior. **Marginal likelihood:**

$$P(\mathrm{x}) = \int (x|\theta)P(\theta)d\theta$$

- **Bayes factor** to compare models:

$$BF_{12}(x) = \frac{P_1(x)}{P_2(x)}$$

- Marginal likelihood is a measure of out-of-sample forecasting performance:

$$P(y_t|y_1, \ldots, y_{t-1}) = \int (y_t|\theta)(\theta|y_1, \ldots, y_t)d\theta$$

- Marginal likelihood is usually very sensitive to the prior.
- **Log Predictive Score** (LPS).

Reference: Mattias Villani (Sveriges Riksbank)

Model Averaging

Collection of models: $M_1, \dots, M_q$.

Posterior model probabilities: $P_{post}(M_i|x) \propto P(x|M_i)P(M_i)$.

Bayesian model averaging: let ξ be any unknown quantity whose interpretation is the same across models:

$$p(\xi) = \sum_{i=1}^{q} P_{post}(M_i|x)p(\xi|M_i)$$

Bayesian prediction (ξ = future value of the process) takes into account

- i) population uncertainty (the error variance)
- ii) parameter uncertainty
- iii) model uncertainty